1. Introduction

Tobacco quality is mainly determined by the maturity stage of the leaves. Only mature leaves show the physical and chemical properties that are well appreciated by smokers and therefore, are requested by the tobacco industry. Only for high quality leaves there is always a demand and an economic profit (Perez-Carbonell, 1987).

There are three main characteristics that define tobacco quality and maturity: *color, texture and aromas*, though color is widely used as the main factor for quality and maturity assessment. However, color varies continuously from green to yellow (commonly called "lemon") and orange within the ripening period, so that commercial color classes show wide fuzzy zones between them.

Color classification is a visual operation, made by experts, and it shows a great level of uncertainty due to the lack of clear edges between the different tobacco quality classes. Moreover, this operation is clearly affected by the expert's fatigue during his work. Within this context the objective for the current research can be summarized as follows:

2. Objective

To establish an objective color classification procedure, capable of avoiding uncertainty in color quality class assessment of tobacco leaves, and that is precise, in the sense of repeatable.

3. Materials

The material used for the current research is tobacco c.v. "Virginia"; this variety covered the 67.7% of the whole Spanish tobacco production in 1993.

The material was provided by CETARSA along the 1994 and 1995 seasons; this company processed 21.000 t of tobacco leaves in 1993 (51.6% of the whole Spanish tobacco production; Anonymous, 1995 a,b). The tobacco samples were previously classified by CETARSA experts in 12 color commercial classes. In 1994 five leaves per class were provided in order to create the classification procedure, while in 1995 two more leaves were included per class in order to validate classification results.

The CETARSA commercial classes are named as shown in Table 1.

4. Methods

As the main objective for the current research was to build up an objective color classifier, the main measurements carried out were color measurements. Color assessment is widely affected by the light used during color observation. In order to avoid the luminiscence variation, the standard method for color observation establishes the use of a 40w daylight lamp or two 36w daylight lamps 84-P30, with a colorimetric index of 80% (F7 Iluminant; Perez-Carbonell, 1987). Therefore this type of illuminant was used for automatic color assessment.

Table 1. Commercial color classes provided by the CETARSA experts in 1994 and 1995, though the "oxidated brown" class was not avaliable in 1995.

Class number	Color	Class number	Color
1	Pale Lemon (PL)	7	Soft Orange (SO)
2	Soft Lemon (SL)	8	Moderate Orange (MO)
3	Moderate Lemon (ML)	9	Hard Orange (HO)
4	Hard Lemon (HL)	10	Deep Orange (DO)
5	Deep Lemon (DL)	11	Light Brown (LB)
6	Pale Orange (PO)	12	Oxidated Brown (OB)

In both seasons, the samples were stored in darkness under controlled temperature and humidity conditions (4° C & 70% R.H.) until they were tested. For data acquisition, the humidity of the tobacco leaves was kept at around 16%, as these conditions make leaves flexible and easy to be handled.

The color tests were carried out using a spectrophotometer (Monolight) enabling to measure the optical reflectance of the visible part of the spectrum. The spectrum was afterwards recalculated by the iluminant characteristics. Integrating these spectra the Triestimulus Color Coordinates: X,Y,Z (equivalent to human perception of color) were calculated (UNE 72-031-83). Also the CIE Coordinates: L,a,b, were used for color assessment (UNE 40-080-84).

For sampling purposes a plastic grid, containing six squared cells, was placed on the face of each extended leave. In each cell two different measurements (observations) were made obtaining a total number of twelve observations per leave, that makes a total number of 780 observations in 1994 and 720 observations in 1995.

5. Data analysis and results

5.1 Consistency of human experts classification between 1994 and 1995

As a first step a study of the average tristimulus values (X,Y,Z) for each CETARSA commercial class was carried out comparing the 1994 and 1995 seasons (see Figure 1). The results obtained for the X and Y values showed that there is a significant decrease of these values, when travelling through Lemon to Orange and Brown classes. The range between "Pale Lemon" and "Light Brown" shows a great similarity of the human classification between seasons, though the range was wider for the 1995 data. From the observation of these data we decided to consider the 1995 range as the correct one. Also, the spread of the mean values of the human expert classes is better in 1995. Therefore, from now on, the human expert class names will be refered to the ones of 1995.

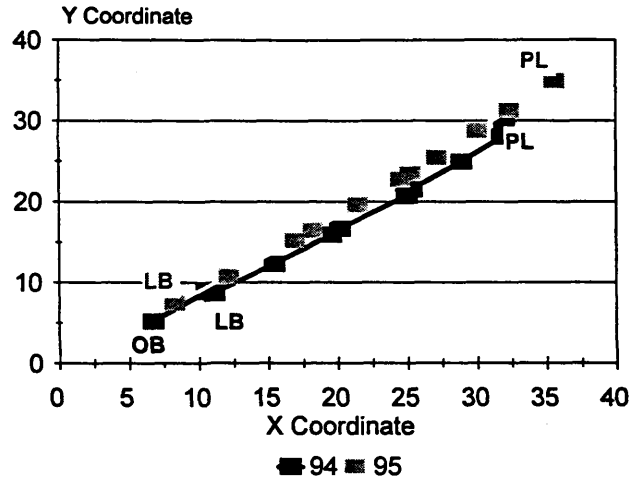

Figure 1. Average values of the X,Y coordinates for each color commercial class ("Pale Lemon, PL, to "Light Brown", LB) shown by the experts classification in 1994 and 1995. A wider color range is appreciated for the 1995 data; note that the "Oxidated Brown" class (OB) was not provided in 1995.

Afterwards, a study of the color variability inside the human experts' classes was studied. A mean coefficient of variation of 20.14% in 1994 and 19.08% in 1995 was observed (see Figure 2, grey points in classes 1 to 12 corresponding to the 1994 data). In this Figure a great overlap between the color commercial classes is shown due to the intragroup variability.

5.2. Generation of a new color database

Due to the lack of well segregated CETARSA classes (original classes) it was decided to employ a non supervised classifier ("clustering" by Ward method; Judez, 1989) to rebuild the color classes. This data procedure means the use of the total number of observations in order to generate new classes through the most similar observations (least distance between observations in the X,Y,Z space); 720 and 660 observations in 1994 and 1995 respectively (as no "oxidated brown" data were provided in the latest season). The observations forming the new classes, "clusters", do not have to belong to the same original leaves, as the system is blind to the initial adscription of the data; the "clustering" method was applied independently for the 1994 and 1995 data in order to establish the robustness of the generation of the new color database.

The number of "clusters" was selected in dependance from the number of classes provided by the human experts (12 in 1994 and 11 in 1995). Table 2 shows that the average Y values for every "cluster" can be directly identified in both series of clusters (1994 and 1995), with a higher Y average in 1995 (as shown in Figure 1). This coincidence between the clustering of both seasons shows the robustness of the new database generation.

When trying to establish a match "cluster"-"human experts' class", the nearest Y values of every "human expert class" for 1995 was used (as decided in paragraph 5.1). This assigment, shown in Table 2, leads to the appearance of a a "void" class of a "Very Pale Lemon" and to an indefinity between "Pale Orange" and "Soft Orange" classes in 1995. Under this criterion no "Pale Lemon" samples were provided in 1994 (already shown in Figure 1).

The clustering method decreases the color variation within the new color classes or "clusters" (from 20.14% to 6.31% in 1994 and from 19.08% to 5.81% in 1995, see black points in Figure 2 corresponding to the 1994 data) obtaining a more homogeneous color data base than the original one, made out of leaves; there is no possibility of obtaining a 100% homogeneous color leaves in nature and so they show a high color variation.Therefore, this new color database offered advantages when comparing with the original one, beeing also season independent.

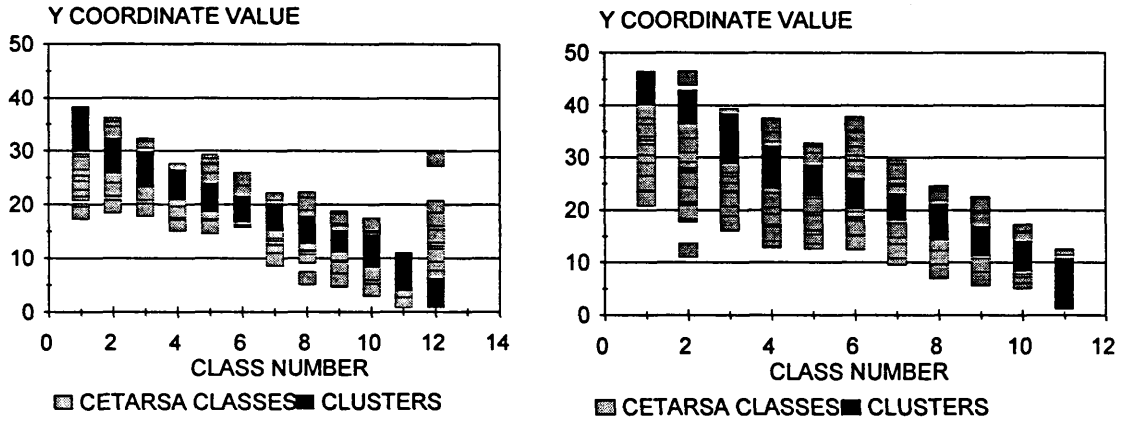

Figure 2. Results obtained with the clustering procedure in 1994 (the grey ponits correspond to the color observations taken on the human experts' classes (60 observations per class). The new color database ("clusters", black points) are formed with the observations which are most similar in the three color coordinates (X,Y,Z). The intragroup Y variance decreases, when compared with the human expert classes (grey points).

5.3. Color classification procedure

The new color classes or "clusters" were used to build up and to test a color classification procedure of the leaves. At this point it was decided to build a color classifier for the observations and to keep for a second step a "total leaf" classification; 720 and 660 observations for the 1994 and 1995 seasons were used, also the validation set in 1995 was tested (2 leaves per color class, 264 observations).

Using the new color database ("clusters"), two different data procedures were used for the classification of the observations: *stepwise discriminant analysis (SDA)*, and *discriminant analysis through Bayes theorem (DAB)*.

Table 2. Average values of variable Y for every cluster in 1994 and 1995. Comparison of "clustering" results for both seasons. The "clusters' " coincidence confirm the lower color range for the 1994 data. The identification between the "clusters" and the experts classes is also indicated.

Clusters 1994		Clusters 1995		Human expert class for 1995
number	Average Y value	number	Average Y value	
		1	43,87	
		2	39,54	Pale Lemon, **PL**
1	32,87	3	33,38	Soft Lemon, **SL**
2	29,54	4	28,9	Moderate Lemon, **ML**
3	26,76	5	26,02	Hard Lemon, **HL**
4	23,63	6	23,1	Deep Lemon, **DL**
5	21,06	7	20,65	Pale Orange, **PO**
6	19,12			Soft Orange, **SO**
7	17,42	8	17,14	Moderate Orange, **MO**
8	15,16	9	14,13	Hard Orange, **HO**
9	13,14			
10	10,95	10	11,47	Deep Orange, **DO**
11	7,09	11	7,49	Light Brown, **LB**
12	4,19			Oxidated Brown, **OB**

Stepwise discriminant analysis (SDA). This analysis procedure generates new variables or "factors" by using the initial parameters (in this case all the mentioned color coordinates: X, Y, Z, x, y, L, a, b) in order to minimize the intragroup variance and maximize the intergroup variance (Judez, 1989). Being a stepwise procedure, the system does not use all the initial variables at the same time. Instead, it selects the most segregating parameter and incorporates afterwards "step by step" other parameters checking whether the class segregation improves or not. In our case, there was only one "discriminant factor" made up by the Y tristimulus value partially (representing green color, related to chlorophyll content). The percentage of well classified observations obtained through this SDA was 91.2.

Discriminant analysis through Bayes theorem (DAB). This type of data analysis does not select the variables by their segregating ability. Therefore, it was decided to use the choosing criterion of the stepwise discriminant analysis. That is, to use the Y tristimulus value as the segregating parameter to be employed.

The Bayes classifier is based on the calculation of the probability of ascription to a certain class or cathegory (SAS, 1988). In order to do so, it is necessary to define the

"specific density functions" for each color class or "cluster". These "density functions" were considered to be Gaussian so that it was possible to create them through the mean and standard desviation of each "cluster".

Once the probability of ascription of each observation to all twelve color classes or "clusters" is made, it is necessary to assign the observation to one of those classes. The criterion used for final class assignation was:

- if the max probability $\geq$ 60% then the class ascription is the one to which the degree of ascription is maximum, and

- if 50% < max probability < 60% then the class ascription is an average value between the class to which the degree of adscription is maximum and the class with a ascription probability inmediatly lower (always an adjoining class).

Within this procedure it is possible to identify those individuals which are between two classes as "frontier observations" (2.6% in 1994 and 4.6% in 1995; see Table 3). The percentage of classification errors obtained using the Bayes classifier was 4.09% in 1994 and 0.57% in 1995. These results improved the stepwise discriminant analysis results, due mainly to the possibility of identifying those "frontier individuals" but also to the use of quadratic instead of linear discriminant functions.

Table 3. Efficiency of the automatic color classifier generated by the stepwise discriminant analysis. Only one discriminant factor was selected by the system, the Y triestimulus value. According to the new color database, no "Pale Lemon" samples were provided in 1994 (already confirmed in Figure 1), Also a single class covering the range "Pale Orange""-"Soft Orange" (See Table 2) was included in 1995 samples.

	Color class as defined in Table 1											
Season	PL	SL	ML	HL	DL	PO	SO	MO	DO	HO	LB	OB
1994	-	95	98	88	100	92	94	100	93	90	87	90
	% TOTAL WELL CLASSIFIED (WC) 93.3% FRONTIER INDIVIDUALS 2.6% ERROR 4.09%											
1995	97	100	94	92	96	97		94	94	96	89	-
	% TOTAL WELL CLASSIFIED (WC) 94.8% FRONTIER INDIVIDUALS 4.6% ERROR 0.57%											

5.4. Average color evaluation of leaves. Comparison with human experts.

At this stage an homogeneous color base as well as a classification procedure was created; the class method enabled to identify the class of ascription for each observation. However, there were twelve observations per leaf so it was necessary to establish an evaluation criterion for the color of each leaf so it could be compared with the initial evaluation of leaves done by the human experts.

An evaluation criterion was selected assigning as final color class the average of the 12 color data taken initially per leaf. The results for color evaluation of leaves in 1995 are shown in Table 4. These results indicate that there is a very high correspondance between the classification done by the human experts and the objective classification. It is evident that there exist some inconsistencies or lack of coherence of the human experts for different seasons.

Table 4. Comparison of the objective optical classifier (columns) with the human experts' (rows). Results for color evaluation of leaves in 1994 and 1995; 5 and 7 leaves per class were provided by the experts in each season respectivelly. To improve the classification complementary information as texture should be included.

	1994												1995									
	PL	SL	ML	HL	DL	PO	SO	MO	HO	DO	LB	OB	PL	SL	ML	HL	DL	PO -SO	MO	HO	DO	LB
PL			4	1									3	3	1							
SL			2	3									1	2	3	1						
ML			1		4										3	4						
HL					1	4										3	4					
DL				1		3	1									1	3	2	1			
PO						4	1								1	1		4	1			
SO							2	3										6	1			
MO							2	3										1	2	4		
HO									4	1									1	6		
DO									1		4									1	5	1
LB									1	2	2											7
OB											2	3										

6. Conclusions

The conclusions obtained in the current research can be summarized as follows:

- the commercial color classes provided by the experts showed a high overlapping, for the tristimulus values, caused by the lack of homogenity in any biological samples; a mean coefficient of variation of 20.14% in 1994 and 19.08% in 1995 was observed in these original classes,

- the clustering method enabled us to create new color classes or "clusters" with the individual observations, which were similar in color to the CETARSA color classes but with less inside-class variability (mean coefficient of variation of 6.31% in 1994 and 5.81% in 1995),

- the Bayes classifier method showed the lowest classification errors (4.09% in 1994 and 0.57% in 1995) due to the use of quadratic discriminant functions, but also to its ability to identify "frontier observations" (2.6% and 4.6% for 1994 and 1995),

- a decision criterion for color ascription of leaves has been developed showing high correspondance with the evaluation of leaves made by the CETARSA experts,

- further research should be carried out in order to combine color with other complementary information like texture, as it seems that human perception does it also unconsciously, and

- all these results will be validated "in situ" during the current season in a collaborative study with CETARSA experts.

7. Acknowledgements

Our gratitude to CETARSA for their collaboration at the current research, and to CICYT, Project n°94-1082: *"Desarrollo de aplicaciones de la reflectancia óptica en las regiones VIS y NIR del espectro para la medida no destructiva de factores de calidad de pimentón con extensión a otros productos agrícolas"*.

8. References

1) Anonymous. 1995 a. *El mercado peninsular de tabaco en 1993*. Tabaco y Noticias n°6 pp:20-23.

2) Anonymous. 1995 b. *Reducción de costes*. Tabaco y Noticias n°7 pp:4-5.

3) Judez L. 1989. *Técnicas de análisis de datos multidimensionales*. Ed. Ministerio de Agricultura, Pesca y Alimentación.

4) Perez Carbonell H. 1987. *El curado del tabaco flue-cured*. Ministerio de Agricultura, Pesca y Alimentación.

5) SAS Institute Inc. 1988. *SAS/STAT* User's Guide, Release 6.03*. Ed. Cany, NC: SAS Institute Inc. 1028 pp

6) UNE 72-031-83. 1983. *Magnitudes colorimétricas*. IRANOR.

7) UNE 40-080-84. 1984. *Determinación de coordenadas cromáticas "CIE"*. IRANOR.